Self-Deception and Choice*

Igor Kopylov *Jawwad Noor*

March 5, 2009

Abstract

While the literature on temptation and self-control has typically viewed the agent as behaving strategically in order to deal with particular desires, we view desires as using self-deception as a strategy to attain fulfillment. Desires motivate the agent to construct rationalizations that justify actions that eventually lead her into temptation: she may form of an exaggerated view of her ability for self-control, or relax her normative attitude toward indulgence. An axiomatic model of temptation-driven self-deception is presented. It is demonstrated that Gul-Pesendorfer's framework puts strong restrictions on the kind of self-deception it can accommodate. Some implications of self-deception for welfare policy are derived.

1 Introduction

"This self-deceit, this fatal weakness of mankind, is the source of half the disorders of human life."

– Adam Smith, The Theory of Moral Sentiments.

The seminal decision-theoretic work on temptation, due to Gul and Pesendorfer [11] (henceforth GP), has given rise to various extensions that incorporate new dimensions of choice under temptation and applications to

*Kopylov is at the Dept of Economics, University of California, Irvine; E-mail: ikopylov@uci.edu. Noor is at the Dept of Economics, Boston University; Email: jnoor@bu.edu.

various topics. Common to all these studies is a particular view of how temptation acts on the agent: desires simply *appear* to the agent at the moment of choice and demand their satisfaction. This paper promotes an alternative view of temptation, one that regards desires as being more strategic and employing more sophisticated means to achieving satisfaction. We hold that, in order to attain satisfaction, desires may motivate a distortion in the agent's perception of herself or of the world in a way that leads her into temptation. In a word, temptation can take a disguised form, one that manifests itself in self-deception.

1.1 Motivation

The psychology literature recognizes the strategic nature of temptation. It studies not only the strategies that people use to resist temptation, but also the strategies they use *to allow their resolve to fail*. Baumeister et al [3, pg 139] write:

> [S]mokers face the choice of obtaining immediate gratification of their addiction or living for a longer period of time. To do this, smokers must ignore or rationalize the long-term consequences of smoking. Thus, for instance, they claim that the evidence is weak linking smoking to cancer, or they fall prey to thoughts that they are personally invulnerable..
>
> Similarly, individuals who may need to lose weight for health reasons often find themselves in tempting situations...Such individuals are known to engage in irrational thought processes during such events ("Well, one cookie won't harm me,"..). Thus, giving in to temptation also involves a number of cognitive strategies that are used to negate the perceived long-term consequences associated with indulgence.

Unsurprisingly, self-deception plays an important role in the literature on addiction. Ludwig [18, pg 12-13] writes:

> [T]he alcoholic's worst enemy is not the bottle or bad luck but his own mind, within which is the ever-present Trojan horse of desire, waiting to smuggle in the enemy when the defenses have been lulled into complacency. What must be recognized is that

> in this case the brain is much less an organ of rationality than of rationalization...
>
> [O]ften, the mind of the abstinent alcoholic is so devious as to come up with a number of seemingly innocuous decisions...that eventually place the individual in a situation in which a return to drink becomes inevitable... [If] he could be really honest with himself, he would recognize that his lapse was not because of bad luck or fate but because, at some level of awareness, he planned to do this all along.

Therefore, desires that directly impact the agent at the moment of choice may also affect the agent in the steps leading up to the moment of choice: they may motivate the agent to construct *rationalizations* for behaviors she hitherto preferred to avoid.[1] These rationalizations enable the agent to justify a course of action that eventually leads her into temptation by diminishing the negative content of the action, often emphasizing the fact that the agent does not perceive or comprehend the 'badness' of her action.[2] Thus, the smoker in the above example explains away the evidence in a self-serving way, and the dieter mentally reframes her problem as one not involving a serious conflict with normative considerations. The facade of reason enables the agent to protect her self-image.

The nature of self-deception, and even its definition, is subject to much philosophical debate. The traditional *intentionalists* understand self-deception in light of interpersonal deception, thus requiring the self-deceiver to believe p but intentionally bring about the belief $\neg p$ (Sartre [22], Fingarette [10]). The difficulties raised by implied paradoxes has led more recent intentionalists to weaken the requirement of holding contradictory beliefs to just intentionally bringing about a belief – motivated by some desire or emotion – despite an initial recognition that the evidence may not warrant this belief (Talbott [24], Bermudez [5]).[3] Levy [17] suggests that the agent can "avoid the evidence, or situations in which he is likely to be confronted by the evidence, can

[1] Kunda [16, pg 482] proposes that people who are "motivated to arrive at a particular conclusion attempt to be rational and to construct a justification of their desired conclusion that would persuade a dispassionate observer. They draw the desired conclusion *only if* they can muster up the evidence necessary to support it" (emphasis added).

[2] These are examples of what Snyder [23] calls a *reframing strategy.*

[3] To be precise, this is a definition of 'straight self-deception'. In cases of 'twisted self-deception' the motivated belief is in fact undesired, such as when a jealous husband believes on weak evidence that his wife is having an affair (Mele).

rationalize the evidence he has by imagining unlikely but possible explanations for each piece, and so on" and Talbott [24] suggests that the agent can exercise selectivity in attention, memory, evidence-gathering and reasoning. In this view, intentionality is not abandoned, and it remains prereflective: the enterprise is prompted by motives that cannot be spelled out without destroying the enterprise itself.[4] *Non-intentionalists* allow that such forms of self-deception may be possible, but emphasize that most cases of self-deception can be also be explained in terms of unintentional motivationally-biased information processing, without resorting to unconscious beliefs and intentions (Barnes [2], Mele [19]). The model of self-deception pursued in this paper is consistent with any of these views.

This paper is concerned primarily with what we refer to as *temptation-driven self-deception*, which is motivated by desires that, according to the agent's normative views, should not be satisfied. In such cases, the agent may distort her perceptions of herself or the world in a way that provides a rationale and preemptive excuse for submitting to her desire, thereby shielding the agent from feelings of guilt, shame or inferiority both in the present and future. For perspective, we mention that an alternative motivation for self-deception may concern long term goals or the need to adjust and function in a social environment. This accommodates the case of *positive illusions* (Taylor [25], Taylor and Brown [27]) that take the form of self-aggrandizing self-perceptions, exaggerated perception of control over surroundings and unrealistic optimism, which are viewed as serving an important adaptive function. Finally, we also mention that *wishful thinking* is related but distinct from self-deception because the latter, prompted by some emotional need, survives in the 'teeth of evidence' (by explaining away or reinterpreting the evidence) while the former does not (Szabados [26]).

1.2 This Paper

We model temptation-driven self-deception in a GP-style setup. Choices are made in three stages—ex ante, interim, and ex post. Consumption takes place only in the ex post stage, where the agent faces a menu to choose out of. This menu is itself chosen in the interim stage, and the menu of menus

[4]Intentionality explains the internal tension typically identified with self-deception, which may be revealed in opacity and indirection (Levy [17]), emotional resistance to evidence (Talbott [24]) or hypersensitivity to criticism, confrontation or opposing evidence (Gur and Sackeim [12]).

available to her in this stage is determined in the ex ante stage. Temptation acts directly on the agent in the ex post stage, and it acts on her indirectly via self-deception in the interim stage. Our primitive is the agent's preferences over menus of menus in the ex ante stage, which is assumed to be prior to the experience of any temptation or self-deception. At this stage, the agent is in a cold state, and anticipates her future experience and struggle with temptation.[5]

In this framework, we introduce and characterize a behavioral axiom of self-deception. As discussed above, self-deception is a strategic manipulation of the agent's reasoning that is motivated by desires. Our axiom draws from this by requiring that the temptation value of a menu is dependent on the extent of ex post satisfaction of desires in that menu. In particular, if temptation is resisted in a menu, the agent is not motivated to choose it in the interim stage. We find that the corresponding representation for preference admits an intuitive interpretation in terms of self-deception.[6] Specifically, the agent behaves as if she is necessarily tempted to *relax her ex ante normative perspective* so as to accommodate a greater degree of indulgence, and may even become *blind to the presence of temptation* in a menu.

Our result conspicuously rules out the possibility that the agent may exaggerate her ability to exert self-control and, more generally, misperceive her future choices. By appealing to behavioral implications of the model, and by constructing a more general model that incorporates virtuous self-perception, we show that a consistent tendency to rationalize by appeal to a virtuous self-perception is inconsistent with self-deception proper: a virtuous self-perception may in fact obstruct, rather than enhance, the satisfaction of desire. The lesson we take away is that the GP axioms put strong restrictions on the nature of self-deception that it can accommodate. For instance, it cannot accommodate the existence of a "local" virtuous self-image, one that arises only when it serves the purpose of desire. A richer model of self-deception would need to be embedded in some extension of GP's model.

The noted general model shares elements with what would be considered a richer model of self-deception – one where the three distinct kinds of dis-

[5]Empirical foundations for such a 'special period 0' perspective are provided in Noor [21] on the basis of the hypothesis that distant temptations have smaller impact on current choices than do immediate temptations. Thus, our ex ante stage may be interpreted as a time sufficiently distant from the interim stage.

[6]As we discuss in Section 3, the 'internal tension' feature is obtained in the GP model by fiat.

tortions in self-perceptions are present. Although it cannot be regarded as a general model of self-deception, it lends itself to a substantive discussion of welfare. For instance, the model implies that agents' welfare is lowered if menus containing temptation ('vice' menus) are mixed with menus containing normatively attractive alternatives ('virtuous' menus). This is reflected in the following preference over menus of menus:

$$\{a, b\} \succeq \{a, a \cup b\},$$

where a is a virtuous menu and b a vice menu. Intuitively, larger menus lend themselves to more excuses. Thus, the agent can more effective deceive herself into choosing $a \cup b$ from $\{a, a \cup b\}$ than choosing b from $\{a, b\}$. This provides a rationale for separating vice and virtue, such as via zoning laws for casinos.

The remainder of this paper proceeds as follows. The introduction concludes with a mention of related literature. Section 2 introduces the primitives of the model and presents a benchmark case. Section 3 presents our model of self-deception. Section 4 extends this model to permit a virtuous self-perception and derives some of its implications for welfare policy. Section 5 concludes. Proofs are relegated to appendices.

1.3 Related Literature

To our knowledge, the notion of temptation-driven self-deception that we focus on has not been studied in economics. The behavioral economics literature discusses positive illusions (Benabou and Tirole [4]) and wishful thinking (Brunnermeier and Parker [6]), and the decision-theoretic literature has modelled temptation only as arising in its naked form and being relevant mainly at the moment of choice.

The idea that temptation may influence the choice of menu is present in Noor [20, 21] but the mode in which temptation acts there is presumed to be direct. That is, a menu tempts the agent just as an alternative in that menu may tempt her. Our paper maintains that an agent may be tempted by menus, but it hypothesizes that the agent is only tempted by menus that she can justify choosing. Indeed, we find that the model of tempting menus in [20] cannot be regarded as one of self-deception.

A version of our self-deception model also shows up in Noor [21]. There are two important differences, however. First, the model appears in [21]

mainly as a means to axiomatically unify other temptation models in the literature. In contrast, the current paper starts by behaviorally defining a necessary condition for self-deception, and it turns out that the representation theorem delivers the same model. A novel interpretation of that model is obtained here as a result. Second, our choice domain differs substantially from [21] and so do our axioms. Intuitively, our axioms are imposed on the agent's perspective in a cold state – in a special period 0 – where she anticipates all temptation (by self-deception or otherwise) but is not subject to it yet. In contrast, axioms in [21] are imposed directly on choice in a hot state, when the agent is subject to all kinds of temptation. A counterpart of our main axiom does not appear in [21].

Finally, on a technical level our results exploit Kopylov's [13, 15] extensions of GP's model to representations with finitely many additive components and to dynamic settings with more than two periods.

2 Preliminaries

To aid exposition in subsequent sections, we first present a basic three-period extension of GP's model.

Let $X = \{x, y, z, \dots\}$ be the set $\Delta(Z)$ of all Borel probability measures on a compact metric space Z of deterministic consumptions. Let d be the Prohorov metric of the weak convergence topology on X. More generally, let X be the class of all Anscombe–Aumann acts f that map a finite state space Ω into $\Delta(Z)$, and let d be the corresponding product metric in X.[7]

Suppose that choices are made in three stages—ex ante, interim, and ex post—and these choices determine the decision maker's consumption in X after the ex post stage. Let $\mathcal{M}_1 = \{a, b, c, \dots\}$ be the set of all interim *menus*—non-empty compact subsets $a \subset X$. Interpret any menu $a \in \mathcal{M}_1$ as a course of action that, if taken at the interim stage, restricts the ex post choice to the set $a \subset X$. Endow the space $\mathcal{M}_1$ with the Hausdorff metric μ_1 and define mixtures

$$\alpha a + (1-\alpha)b = \{\alpha x + (1-\alpha)y \ : \ x \in a, \ y \in b\}$$

for all $\alpha \in [0,1]$ and menus $a, b \in \mathcal{M}_1$.

[7]Menus of lotteries are first used by Gul and Pesendorfer [11] and—for finite Z—by Dekel, Lipman, and Rustichini [7]. Menus of acts are proposed by Epstein [9].

Similarly, let $\mathcal{M}_0 = \{A, B, C, \dots\}$ be the set of all ex ante menus—non-empty compact subsets $A \subset \mathcal{M}_1$. Interpret any menu $A \in \mathcal{M}_0$ as a course of action that, if taken ex ante, restricts the interim choice to the set $A \subset \mathcal{M}_1$. Endow the space $\mathcal{M}_0$ with the Hausdorff metric μ_0 and define mixtures

$$\alpha A + (1-\alpha)B = \{\alpha a + (1-\alpha)b \,:\, a \in A,\ b \in B\}$$

for all $\alpha \in [0,1]$ and menus $A, B \in \mathcal{M}_0$. Then both $\mathcal{M}_0$ and $\mathcal{M}_1$ are compact (see Theorem 3.71 in Aliprantis and Border [1]) and the mixture operations in these spaces are continuous.

Let a binary relation $\succeq$ on $\mathcal{M}_0$ be the decision maker's weak preference over ex ante menus. Write the symmetric and asymmetric parts of this relation as $\sim$ and $\succ$ respectively. Note that our model does not take the decision maker's interim and ex post choices as primitive, but instead derives her anticipation of these choices from her ex ante preference.

Adapt GP's list of axioms for the preference $\succeq$.

Axiom 1 (Order). *$\succeq$ is complete and transitive.*

Axiom 2 (Continuity). *For all menus $A \in \mathcal{M}_0$, the sets $\{B \in \mathcal{M}_0 : B \succeq A\}$ and $\{B \in \mathcal{M}_0 : B \preceq A\}$ are closed.*

Axiom 3 (Independence). *For all $\alpha \in [0,1]$ and menus $A, B, C \in \mathcal{M}_0$,*

$$A \succeq B \quad \Rightarrow \quad \alpha A + (1-\alpha)C \succeq \alpha B + (1-\alpha)C.$$

Axiom 4 (Set-Betweenness). *For all menus $a, b \in \mathcal{M}_1$ and $A, B \in \mathcal{M}_0$,*

$$\{a\} \succeq \{b\} \quad \Rightarrow \quad \{a\} \succeq \{a \cup b\} \succeq \{b\}, \tag{1}$$
$$A \succeq B \quad \Rightarrow \quad A \succeq A \cup B \succeq B. \tag{2}$$

Order and Continuity are standard conditions of rationality. To motivate Independence, interpret any mixture $\alpha A + (1-\alpha)C$ as a lottery that yields the menus A and C with probabilities α and $1-\alpha$ respectively and is resolved after the ex post stage. In this interpretation, the decision maker's interim choice $\alpha a + (1-\alpha)c$ in $\alpha A + (1-\alpha)C$ and her ex post choice $\alpha x + (1-\alpha)y$ in $\alpha a + (1-\alpha)c$ determine her consumptions $x \in a \in A$ and $y \in c \in C$ contingent on the resolution of the lottery between the menus A and C. If the timing of the resolution of this objective uncertainty is irrelevant for preference, then the decision maker should be indifferent between the menu

$\alpha A + (1-\alpha)C$ and a hypothetical lottery $\alpha \circ A + (1-\alpha) \circ C$ that yields the menus A or C with probabilities α and $1-\alpha$ respectively, but is resolved immediately after the ex ante stage. (Here the preference $\succeq$ is extended from the original domain $\mathcal{M}_0$ to lotteries over menus.) Then the standard separability argument suggests that

$$A \succeq B \quad \Rightarrow \quad \alpha \circ A + (1-\alpha) \circ C \succeq \alpha \circ B + (1-\alpha) \circ C$$

because the possibility of getting the menu C with probability $1-\alpha$ should not affect the decision maker's comparison of A and B. Independence follows.

Set-Betweenness is imposed separately on the preference $\succeq$ over the entire $\mathcal{M}_0$ and on the restriction of $\succeq$ to singleton menus $\{a\}$ that provide a strict commitment to the menu $a \in \mathcal{M}_1$ at the interim stage, but may still require self-control ex post when choice in a has to be made. It is assumed that the decision maker's ex ante evaluation of any such menu $\{a\}$ is based on her anticipation of two factors:

- the consumption $x_a \in a$ that she will choose if a is feasible ex post,
- the self-control that she will use to resist the strongest temptation $y_a \in a$ in this menu.

This informal assumption suggests that for all menus $a, b \in \mathcal{M}_1$,

$$x_b \in a,\ y_a \in b \quad \Rightarrow \quad \{a\} \succeq \{b\}. \tag{3}$$

Indeed, if $x_b \in a$ and $y_a \in b$, then the decision maker should expect that if she chooses x_a from the menu a at the ex post stage, then she will (i) obtain the same consumption that she plans to choose from b and (ii) resist the temptation y_a, which belongs to b and hence, should not be harder to resist than the *strongest* temptation y_b in b. Therefore, the ranking $\{a\} \succeq \{b\}$ is intuitive because the menu a offers a weakly better combination of consumption benefits and self-control costs than b does. Condition (3) implies (1).[8] Analogously, condition (2) assumes that the decision maker should evaluate any menu $A \in \mathcal{M}_0$ based on her anticipated interim choice $a_A \in A$ and the most tempting alternative $b_A \in A$ in this menu. Note that if

[8]To show this claim, take any menus $\{a\} \succeq \{b\}$. Let $c = a \cup b$. Then $x_a, x_b, y_a, y_b \in c$. By (3), if $x_c \in a$, then $\{a\} \succeq \{c\}$; if $x_c \in b$, then $\{b\} \succeq \{c\}$. In either case, $\{a\} \succeq \{c\}$. By (3), if $y_c \in a$, then $\{c\} \succeq \{a\}$; if $y_c \in b$, then $\{c\} \succeq \{b\}$. In either case, $\{c\} \succeq \{b\}$.

temptations are cumulative or uncertain (as in Dekel, Lipman, Rustichini [8]), then both parts of Set-Betweenness can be violated.

The following condition is used to obtain uniqueness in representation results below.

Axiom 5 (Self-Control). *There are $x, y, x', y' \in X$ such that*

$$\{\{x\}\} \succ \{\{x, y\}\} \succ \{\{y\}\}$$
$$\{\{x'\}\} \succ \{\{x'\}, \{y'\}\} \succ \{\{y'\}\}.$$

This axiom requires that the decision maker should expect to have some self-control at the ex post and interim stages, and plan to choose the alternative x and the singleton menu $\{x'\}$ rather than the more tempting y and $\{y'\}$ respectively.

Say that a function $u : X \to \mathbb{R}$ is *linear* if for all $\alpha \in [0, 1]$ and $x, y \in X$,

$$u(\alpha x + (1 - \alpha)y) = \alpha u(x) + (1 - \alpha)u(y).$$

Let $\mathcal{U}$ be the set of all continuous linear functions $u : X \to \mathbb{R}$. Similarly, define linearity for functions on $\mathcal{M}_1$ and let $\mathcal{U}_1$ be the set of all continuous linear functions $V : \mathcal{M}_1 \to \mathbb{R}$.

Theorem 1. *The preference $\succeq$ satisfies Axioms 1–4 if and only if $\succeq$ is represented by a utility function U_0 such that for all $A \in \mathcal{M}_0$ and $a \in \mathcal{M}_1$,*

$$U_0(A) = \max_{a \in A} \left[U(a) - \max_{b \in A} (V(b) - V(a)) \right] \tag{4}$$
$$U(a) = \max_{x \in a} [u(x) - \max_{y \in a} (v(y) - v(x))], \tag{5}$$

where $u, v \in \mathcal{U}$ and $V \in \mathcal{U}_1$.

Moreover, if $\succeq$ satisfies Self-Control, then it has another representation (4) with components $u', v' \in \mathcal{U}$ and $V' \in \mathcal{U}_1$ if and only if $u' = \alpha u + \beta_u$, $v' = \alpha v + \beta_v$, and $V' = \alpha V + \beta_V$ for some $\alpha > 0$ and $\beta_u, \beta_v, \beta_V \in \mathbb{R}$.

This theorem provides a joint characterization for GP's utility representations (4) and (5) over $\mathcal{M}_0$ and $\mathcal{M}_1$ respectively. The restriction of U to $\mathcal{M}_1$ can be interpreted as the decision maker's ex ante normative perspective on what menu *should* be chosen at the interim stage. Temptations that impact her at this stage are captured by V, and the nonnegative component

$\max_{b \in A}(V(b) - V(a))$ is interpreted as the self-control cost of choosing a from A. Similarly, the function u can be interpreted as the ex ante normative perspective on what should be chosen at the ex post stage, and the nonnegative term $\max_{y \in a}(v(y) - v(x))$ as the mental cost of ex post self-control. These interpretations suggest that in order to balance her normative perspective with the costs of self-control, the decision maker should plan to maximize $U + V$ and $u + v$ respectively at the interim and ex post stages.

Before turning to our models of distorted self-perception, note the benchmark case with $V = 0$ when the decision maker does not expect to have any temptations at the interim stage. In this case, she obeys *strategic rationality* so that for all $A, B \in \mathcal{M}_0$

$$A \succeq B \quad \Rightarrow \quad A \sim A \cup B.$$

Note that she may still anticipate costly temptations at the ex post stage, as she may exhibit a preference for commitment $\{a\} \succ \{a \cup b\}$ for some menus $a, b \in \mathcal{M}_1$.

3 Self-Deception

We model *self-deception* as an interim temptation. The ex ante perspective is the unmotivated perspective, but in the interim stage the agent is motivated to change her perspective. Her vehicle for doing so is rationalizations, based on the (unmodelled) process of selectively accessing her memory or gathering evidence and reinterpreting or explaining away evidence. If the motivation to change her perspective is strong enough, she fails to recognize her attempts at self-deception, and interim choice follows her motivated perspective. Intermediate levels of motivation are accompanied at best with *suspicion* of her self-deception, but she is never clear-eyed. Greater degrees of suspicion lead to greater underweighting of the new conclusions she is driven to draw on the basis of the interim evaluation of her evidence, and this underweighting is associated with a cost of disowning the very attractive reasoning.

Suspicion of self-deception plays a role here analogous to 'self-control' in GP's model. However it should be observed that the latter notion cannot be invoked here without giving rise to a contradiction: exertion of self-control against self-deception presupposes knowledge of self-deception, *but there can be no self-deception to begin with if there is knowledge of self-deception.*

3.1 Axiom

An assessment our axiom for self-deception requires us to first identify some possible rationalizations an agent may invoke in the interim period to manipulate herself into temptation. Consider the following three rationalizations:

- Overestimation of propensity for self-control: "I will be tempted, but I can stop myself after one drink, so there's no much harm hitting the bar tonight".
- Underestimation of susceptibility to temptation: "I am not even going to be tempted to have more than one drink".
- Relaxation in normative standards: "I am only going to live once, so I really *should* allow myself to enjoy life a little more".

Each of these rationalizations may be used to justify a choice of a menu that leads the agent into temptation. They diminish the negative value of making such a choice by either denying the existence of temptation or by acknowledging the temptation but denying the possibility of a normatively-bad outcome.

Our axiom for self-deception reflects the *motivated* nature self-deception. It assumes roughly that the agent cannot be motivated (or at least not strongly motivated) to choose a menu that fails to lead to the satisfaction of her temptation preference.

Axiom 6 (Self-Deception). *For all* $a, b \in \mathcal{M}_1$*, if* $\{a \cup b\} \succ \{b\}$*, then*

$$\{a\} \sim \{a, a \cup b\} \quad \text{and} \quad \{b, a \cup b\} \succ \{b\}.$$

The ranking $\{a \cup b\} \succ \{b\}$ suggests that the anticipated ex post choice in the menu $a \cup b$ belongs to a rather than to b. In the case where $\{a\} \sim \{a \cup b\}$ (i.e., b does not contain greater temptation than a) the axiom is readily defensible. The conclusion that $a \cup b$ does not tempt a (i.e. $\{a\} \sim \{a, a \cup b\}$) holds since both menus contain the same temptation and the same final choice. The conclusion that b is not chosen over $a \cup b$ (i.e., $\{b, a \cup b\} \succ \{b\}$) holds since there is neither a normative or temptation desire to do so.

Consider next the nontrivial case where $\{a\} \succ \{a \cup b\}$, that is, where b contains greater temptation than a. Since a and $a \cup b$ lead to the same final choice, and thus the same degree of satisfaction of desires, we hold that there

can simply be no *motivation* for the agent to deceive herself into choosing $a \cup b$ over a. Responding to the extra temptations in $a \cup b$ has no strategic value, and thus cannot motivate the agent to distort her perceptions of her ex post choice or temptation – as in the noted three rationalizations – in order to induce a desire for $a \cup b$ over a. Thus $\{a\} \sim \{a, a \cup b\}$. We hold also that the agent should not be able to deceive herself into choosing b rather than $a \cup b$. Observe that distortions in anticipated choice based on the three rationalizations cannot induce even a strict *desire* for b over $a \cup b$, because $a \cup b$ contains anything that can be chosen in b. Thus, in such a case, we have $\{a \cup b\} \sim \{b, a \cup b\}$, which in turn implies $\{b, a \cup b\} \succ \{b\}$.

One case where a strict desire for b may arise is if the agent views her actual ex post choice as her temptation. For instance, given the preference $\{a \cup b\} \succ \{b\}$ hypothesized in the axiom, the agent may completely reverse her normative and temptation perspectives and say: "I should indulge completely in the menu $a \cup b$, but I will be *tempted to resist*". Thus, if the agent manipulates herself into believing that her perceived choice (based on any of the three rationalization) out of $a \cup b$ lies in b, she may perceive this choice as involving a struggle with a temptation to choose an alternative that lies in a. In such a case b is strictly more desirable than $a \cup b$ according to the distorted perspective. It is not obvious that such cases are those of self-deception: such extreme distorted perceptions appear characteristic of an agent who has completely submitted to temptation rather than one who is delicately trying to get past her normative defenses by appealing to reason. Nevertheless, it should be noted that our axiom does not rule out such possibilities since it permits $\{a \cup b\} \succ \{b, a \cup b\}$. However, it disallows the appeal of such interim temptation to exceed the direct appeal of temptations in b at the ex post stage: if $\{a \cup b\} \succ \{b\}$, so that the temptation in b is not strong enough to impact choice in $a \cup b$, then our axiom requires that a motivated interim temptation by b is also not strong enough to impact choice in $\{b, a \cup b\}$.

3.2 Representation Result

Say that functions $u, v \in \mathcal{U}$ are *independent* if for all $\alpha, \beta, \gamma \in \mathbb{R}$,

$$\alpha u + \beta v + \gamma = 0 \quad \Rightarrow \quad \alpha = \beta = \gamma = 0.$$

Note that u and v are independent if and only if the functions $u, v, u + v$ represent three different rankings on X.

Theorem 2. *$\succeq$ satisfies Axioms 1–6 if and only if $\succeq$ has a utility representation* (4)–(5) *such that for all $a \in \mathcal{M}_1$,*

$$V(a) = \kappa U(a) + \lambda \max_{y \in a} v(y), \tag{6}$$

where $\kappa \geq \lambda > 0$, and $u, v \in \mathcal{U}$ are independent.

Moreover, $\succeq$ has another representation (4), (5), (6) *with parameters $\kappa', \lambda' \in \mathbb{R}$ and functions $u', v' \in \mathcal{U}$ if and only if $\kappa' = \kappa$, $\lambda' = \lambda$, $u' = \alpha u + \beta_u$, and $v' = \alpha v + \beta_v$ for some $\alpha > 0$ and $\beta_u, \beta_v \in \mathbb{R}$.*

Here the temptation utility V can be interpreted as a distortion of the interim normative utility U in the direction of the ex post desires v. This distortion takes the form

$$V(a) = (\kappa - \lambda) \max_{x \in a} \left[\tfrac{\kappa}{\kappa - \lambda} u(x) + \tfrac{\lambda}{\kappa - \lambda} v(x) - \max_{y \in a} (v(y) - v(x)) \right] \tag{7}$$

if $\kappa > \lambda$, or

$$V(a) = \kappa \max_{x \in a} (u(x) + v(x)) \tag{8}$$

if $\kappa = \lambda$. To interpret, compare (7) with (5) to see that interim desires V in (7) modify the ex ante perspective U by replacing the ex ante normative perspective u with

$$u^* = \tfrac{\kappa}{\kappa - \lambda} u + \tfrac{\lambda}{\kappa - \lambda} v.$$

The perspective underlying u^* distorts u in the direction of the temptation utility v. It is as if the decision maker is tempted to believe that her ex ante normative perspective was too stoic, and that she *should* permit herself to follow her desires to a greater extent according to a new normative perspective u^* that more closely follows v. The case (8) is a limiting case of (7) where the decision maker is tempted to view $u + v$ as her normative perspective and to turn a blind eye to any possibility of temptation. It is as if by adjusting her normative perspective she believes that she is resolving all internal conflict.

A surprising observation is that, according to the interim temptation perspective in both (7) or (8), the agent's anticipated ex post choice maximizes $u + v$. That is, anticipated choice is undistorted. Therefore, while representations (7) and (8) accommodate a distortion in normative perspective and possible blindness to temptation, *they do not accommodate a distortion in anticipated self-control ability.* Evidently then, of the three rationalizations

our axiom is consistent with, the third must always hold and the second may hold simultaneously, but the first can never hold. Moreover, anticipated choice is never distorted by any rationalization she adopts.

These conclusions are not tied to the functional form, but rather are supported with behavior. The statement that our self-deceived agent is necessarily tempted to change her normative perspective is captured in the fact that there must exist *singleton* menus $a, b \in \mathcal{M}_1$ such that $\{a\} \succ \{a, b\}$. Since the evaluation of singletons does not involve any (non-trivial) evaluation of ex post choice, this expresses an interim conflict surrounding only what *should* be consumed ex post. The statement that our self-deceived agent is not tempted to misperceive ex post choice (and thus self-control ability) is reflected in the following behavior:[9]

$$\{a\} \succ \{a \cup b\} \Longrightarrow \{a, b\} \sim \{a, a \cup b\}. \tag{9}$$

That is, if b contains tempting alternatives, then there is never a situation where the temptation by $a \cup b$ differs from the temptation by b. A difference would arise if, for instance, the agent's actual anticipated choice from $a \cup b$ was in b (ex ante anticipated lack of self-control) but she was tempted to misperceived her choice from $a \cup b$ to lie in a (temptation-anticipated exertion of self-control). In such a case the agent would be tempted to view $a \cup b$ more favorably than b, and consequently, $\{a, b\} \succ \{a, a \cup b\}$.

4 Multiple Self-Deceptions

Our self-deception model rules out the possibility of a virtuous distortion in her perceived ability to exert self-control. We present here an extension that accommodates such virtuous distortion of self-perception. Although the general model will not be compatible with an interpretation involving sophisticated desires with a single motive, it helps identify some intuitive behaviors ruled out by the model in the previous section and lends itself to discussion of welfare.

[9] To show this claim, take any $a, b \in \mathcal{M}_1$ such that $\{a\} \succ \{a \cup b\}$. Then $V(a \cup b) \geq V(b)$ because $U(a \cup b) \geq U(b)$ and $\max_{y \in a \cup b} v(y) = \max_{y \in b} v(y)$. Consider two cases.

(i) $\{a\} \sim \{a, a \cup b\} \succ \{a \cup b\}$. Then $V(a) \geq V(a \cup b) \geq V(b)$ and hence, $\{a\} \sim \{a, b\}$.

(ii) $\{a\} \succ \{a, a \cup b\}$. By Self-Deception, $\{a \cup b\} \sim \{b\}$, that is, $U(a \cup b) = U(b)$. By (6), $V(a \cup b) = V(b)$. Thus, $\{a, b\} \sim \{a, a \cup b\}$.

In the self-deception model, temptation by b is motivated by the choice that the agents expects to make in b. The following axiom accommodates temptations that are motivated also by the normative content in b.

Axiom 7 (Binary Self-Deception). *For all $a, b \in \mathcal{M}_1$,*

(i) if $\{a \cup b\} \succ \{b\}$, then $\{b, a \cup b\} \succ \{b\}$,

(ii) if $\{a \cup b\} \succ \{b\}$ and $\{a\} \succ \{a, a \cup b\}$, then there is $z \in b$ such that $\{\{z\}\} \succ \{\{x\}\}$ for all $x \in a$.

This condition relaxes Self-Deception and allows the menu $a \cup b$ to tempt a when the anticipated ex post choices in both menus are the same, but the menu $a \cup b$ provides a more virtuous interim self-perception. Formally, the rankings $\{a \cup b\} \succ \{b\}$ and $\{a\} \succ \{a, a \cup b\}$ are allowed only if $a \cup b$ has an element z that is normatively better than any alternative in a. It is as if the agent has an exaggerated view of her propensity for self-control.

A departure from a single-motive self-deception is evident here: interim temptation is no longer intimately connected with a desire to achieve ex post satisfaction of desires. An excessively virtuous self-image may lead the agent to temptation only by accident, and in some situations may even *defeat* efforts at ex post desire satisfaction. It follows that the tendency toward a virtuous self-perception satisfies a different desire – presumably the desire for a positive self-image – that is distinct from ex post temptation. We thus interpret the axiom as adding a second kind of self-deception, one involved in *positive illusions*. However, due to the abstract nature of our framework, we cannot strictly speaking justify this interpretation relative to, say, wishful thinking.

Theorem 3. *$\succeq$ satisfies Axioms 1–5 and Binary Self-Deception if and only if $\succeq$ has a utility representation (4)–(5) such that for all $a \in \mathcal{M}_1$,*

$$V(a) = \kappa U(a) + \lambda \max_{y \in a} v(y) + \mu \max_{z \in a} u(z) \tag{10}$$

where $\kappa \geq \lambda > 0$, $\mu \geq 0$, and $u, v \in \mathcal{U}$ are independent.

Moreover, $\succeq$ has another representation (4), (5), (6) with parameters $\kappa', \lambda', \mu' \in \mathbb{R}$ and functions $u', v' \in \mathcal{U}$ if and only if $\kappa' = \kappa$, $\lambda' = \lambda$, $\mu = \mu'$, $u' = \alpha u + \beta_u$, and $v' = \alpha v + \beta_v$ for some $\alpha > 0$ and $\beta_u, \beta_v \in \mathbb{R}$.

The difference between (10) and (6) is the additional term $\mu \max_{z \in a} u(z)$ in the interim temptation utility V. Although the model offers up to three rationalizations for choosing a given menu (namely, distorted normative preference, distorted temptation preference and perceived virtuosity), the rationalizations may be *inconsistent*. For instance, if the menu contains irresistible temptation, then one rationalization will recognize this and justify it according to a more relaxed normative perspective, while the other will refuse to recognize it.[10] The rationalizations may also *neutralize* each other, such as when a is more virtuous than b and b contains greater temptation. These are reflections of the fact noted above that the model is not one of temptation-driven self-deception. Such self-deception would presumably involve a search for the strongest possible case for making a 'bad' decision, and such a case would be as devoid of inconsistencies as possible. Nevertheless, as a model of an agent who engages in rationalizations more broadly, it permits some substantive discussion of welfare, to which we now turn.

The implication (9) of the pure self-deception model does not hold in general. Formally, write $a \succeq_0 b$ if there exists $z \in a$ such that $\{\{z\}\} \succeq \{\{x\}\}$ for all $x \in b$. Then the preference $\succeq$ represented by (10) satisfies[11]

$$\{a\} \succ \{a \cup b\} \text{ and } a \succeq_0 b \quad \Rightarrow \quad \{a, b\} \succeq \{a, a \cup b\}.$$

This condition implies that if b is a 'vice' menu and a is a 'virtuous' menu, then the agent is generally better off if virtue and vice are kept separate (as in $\{a, b\}$) rather than combined (as in $\{a, a \cup b\}$). The intuition is that the union $a \cup b$ lends itself to *more* excuses for the agent to lead herself into temptation. Indeed, in this case, given whatever rationalizations the agent may adopt to justify choosing b from $\{a, b\}$, virtuous self-perception is an additional rationalization that can be invoked to justify choosing $a \cup$

[10] Behaviorally, suppose $\{\{x\}\} \succ \{\{x, y\}\} \sim \{\{y\}\}$. Then $\{x, y\}$ may be more tempting than both $\{x\}$ and $\{y\}$. That is, $\{\{x\}, \{y\}\} \succ \{\{x\}, \{y\}, \{x, y\}\}$. Since one rationalization favors $\{x\}$ and the other favors $\{y\}$, the fact that $\{x, y\}$ is more tempting than either implies the simultaneous use of both rationalizations.

[11] To show this claim, take any $a, b \in \mathcal{M}_1$ such that $\{a\} \succ \{a \cup b\}$ and $\max_{z \in a} u(z) \geq \max_{z \in b} u(z)$. Then $V(a \cup b) \geq V(b)$ because $U(a \cup b) \geq U(b)$ and $\max_{y \in a \cup b} v(y) = \max_{y \in b} v(y)$. Consider two cases.

(i) $\{a\} \sim \{a, a \cup b\}$. Then $V(a) \geq V(a \cup b) \geq V(b)$ and hence, $\{a\} \sim \{a, b\}$.

(ii) $\{a\} \succ \{a, a \cup b\}$. By Binary Self-Deception, $\{a \cup b\} \sim \{b\}$, that is, $U(a \cup b) = U(b)$. Therefore, $V(a \cup b) \geq V(b)$ implies $\{a, b\} \succeq \{a, a \cup b\}$.

b from $\{a, a \cup b\}$. The behavioral implication suggests, for instance, that a procrastinor is better-off if a completely flexible option is not a feasible choice: if a is the option of completing the task sooner and b is the option of completing it later, then having the opportunity to make this decision later (as in $\{a, a \cup b\}$) makes her worse-off relative to a situation where she has to decide today whether to complete the task sooner or later (as in $\{a, b\}$). For another example, view a menu as a physical location selling particular alternatives and a menu of menus as a town. Then agents in a town are better-off if virtue and vice are sold at distinct locations (as in $\{a, b\}$) relative to when vice is always bundled with virtue (as in $\{a, a \cup b\}$). Zoning laws for casinos may be welfare improving in this sense.

A common view is that optimal welfare policy for agents with self-control problems constitutes the provision of commitment opportunities. In the above setting, this would correspond to providing the agent with $\{a, a \cup b\}$, in which she may keep all her options by selecting $a \cup b$ or avoid temptation by choosing the commitment option a. The above discussion suggests that when agents are subject to self-deception, then the simple provision of commitment opportunities may not always be optimal.

5 Comparative Self-Deception

To interpret the parameters κ, λ, and μ in terms of choice behavior, consider a pair of preferences $\succeq$ and $\succeq^*$ over $\mathcal{M}_0$. Call this pair *regular* if both $\succeq$ and $\succeq^*$ satisfy Axioms 1–5 and BSD, and the two rankings agree on the domain of singleton menus so that

$$\{a\} \succeq \{b\} \quad \Leftrightarrow \quad \{a\} \succeq^* \{b\}$$

for all $a, b \in \mathcal{M}_1$.

By Theorem 3, any regular pair of preferences $\succeq$ and $\succeq^*$ can be represented by (4)-(6) with components $(u, v, \kappa, \lambda, \mu)$ and $(u^*, v^*, \kappa^*, \lambda^*, \mu^*)$ respectively. Moreover, the functions U and U^* represent the same preference on $\mathcal{M}_1$ and hence, by GP's Theorem, one can take $u = u^*$ and $v = v^*$.

Say that $\succeq^*$ is *more self-deceptive* than $\succeq$ if for all menus $a, b \in \mathcal{M}_1$,

$$\{a\} \succ \{a, b\} \quad \Rightarrow \quad \{a\} \succ^* \{a, b\}, \tag{11}$$

This definition requires that any self-deception that is tempting for $\succeq$ should be tempting for $\succeq^*$ as well.

Theorem 4. *Let $\succeq$ and $\succeq^*$ be a regular pair of preferences. Then $\succeq^*$ is more self-deceptive than $\succeq$ if and only if the two preferences have representations (4)-(6) such that $\frac{\lambda^*}{\kappa^*} \geq \frac{\lambda}{\kappa}$ and $\frac{\mu^*}{\lambda^*} = \frac{\mu}{\lambda}$.*

This result suggests that the ratios $\frac{\lambda}{\kappa}$ and $\frac{\mu}{\kappa}$ are both positively related to the intensity of self-deception in our model. (Note that $\frac{\lambda^*}{\kappa^*} \geq \frac{\lambda}{\kappa}$ and $\frac{\mu^*}{\lambda^*} = \frac{\mu}{\lambda}$ imply $\frac{\mu^*}{\kappa^*} \geq \frac{\mu}{\kappa}$.) If $\kappa > \lambda$, then the weight $\frac{\lambda}{\kappa-\lambda}$ that is put on the function v in (7) is a positive monotonic transformation of $\frac{\lambda}{\kappa}$ and hence, can serve as an index of self-deception as well.

Moreover, the equality $\frac{\mu^*}{\lambda^*} = \frac{\mu}{\lambda}$ is necessary for an unambiguous comparison of self-deception revealed by the two rankings $\succeq$ and $\succeq^*$. This equality requires roughly that the proportion of the virtuous and motivated components of self-deception should be the same for $\succeq$ and $\succeq^*$. In particular, this must be true when both $\succeq$ and $\succeq^*$ satisfy Self-Deception and hence, $\mu = \mu^* = 0$.

6 Conclusion

On an intuitive level, one would expect that an exaggerated self-perception of virtuosity may serve as a strategic tool for desires to induce the agent to make decisions that will lead her into temptation. Yet, necessary behavioral properties of self-deception (Motivated Self-Deception and Self-Deception), when imposed on a GP-style model, do not permit any such distorted self-perceptions and instead necessitate a distortion in one's normative perspective and may imply complete denial regarding the possibility of temptation against the new normative perspective. In particular, any distortion in expectation about future choice behavior (and propensity for self-control) is ruled out. Thus, we find that in order to accommodate the intuitive possibility of a distortion in perception of future choice behavior to be 'triggered' whenever it serves desires, one needs to go beyond GP's model.

We envision a general model of self-deception to have the following features. There may be a set of rationalizations that the agent can adopt, which may include distorted normative perspectives, virtuous self-perception, or a biased interpretation of evidence (not modelled in this paper). At the time of interim choice, particular rationalizations are adopted *only* if they serve desire, that is, only when they cause the agent to choose a menu containing irresistible temptation. How tempting a menu is at the interim stage de-

pends on how many rationalizations the agent can conjur up and whether these rationalizations are convincing in the sense of not contradicting each other. Indeed, the existence of consistent rationalizations may be a necessary condition for a menu to even tempt.

A APPENDIX: PROOFS

In the proofs, we use the following notation and terminology. For any function $u \in \mathcal{U}$ and any menu $a \in \mathcal{M}_1$, write

$$u(a) = \max_{x \in a} u(x),$$

and let

$$\mathcal{T}(u) = \{\alpha u + \beta : \alpha \geq 0, \beta \in \mathbb{R}\}$$

be the set of all non-negative transformations of the function u. Say that functions $u_1, \ldots, u_n \in \mathcal{U}$ are *redundant* if there is a constant function u_i in this list, or if $u_i = \kappa u_j + \beta$ for some $i \neq j$, $\kappa > 0$ and $\beta \in \mathbb{R}$.

For any $S \in \mathbb{N}$, let S denote also the set $\{1, \ldots, S\}$. The following result is invoked from Kopylov [13].

Lemma 5. *For any $u_1, \ldots, u_S \in \mathcal{U}$, there are elements $x_1, \ldots, x_S \in X$ such that for all $i, j \in S$, $u_i(x_i) \geq u_i(x_j)$, and*

$$u_i \notin \mathcal{T}(u_j) \quad \Leftrightarrow \quad u_i(x_i) > u_i(x_j). \tag{12}$$

Proof. Take any $k, l \in S$ and consider two possible cases.

(i) $u_k \in \mathcal{T}(u_l)$. Take $x_{kl}, y_{kl} \in X$ such that $u_k(x_{kl}) \geq u_k(y_{kl})$.

(ii) $u_k \notin \mathcal{T}(u_l)$. Then by Herstein–Milnor's Theorem, there are $x_{kl}, y_{kl} \in X$ such that $u_k(x_{kl}) > u_k(y_{kl})$ and $u_l(y_{kl}) \geq u_l(x_{kl})$.

For all $i \in S$, let $x^i_{kl} = x_{kl}$ if $u_i(x_{kl}) > u_i(y_{kl})$ and $x^i_{kl} = y_{kl}$ otherwise. Let

$$x_i = \sum_{k,l \in S} \tfrac{1}{S^2} x^i_{kl}.$$

Take any $i, j \in S$. Then $u_i(x^i_{kl}) \geq u_i(x^j_{kl})$ for all $k, l \in S$, and hence,

$$u_i(x_i) = \sum_{k,l \in S} \tfrac{1}{S^2} u_i(x^i_{kl}) \geq \sum_{k,l \in S} \tfrac{1}{S^2} u_i(x^j_{kl}) = u_i(x_j).$$

Moreover, if $u_i \notin \mathcal{T}(u_j)$, then

$$u_i(x_{ij}^i) = u_i(x_{ij}) > u_i(y_{ij}) = u_i(x_{ij}^j),$$

and hence, $u_i(x_i) > u_i(x_j)$. Conversely, the inequalities $u_i(x_i) > u_i(x_j)$ and $u_j(x_j) \geq u_j(x_i)$ imply that $u_i \notin \mathcal{T}(u_j)$. □

The necessity of Axioms 1–4 in Theorem 1 is straightforward. Conversely, suppose that $\succeq$ satisfies Axioms 1–4. Kopylov [15, Theorem 1] shows that $\succeq$ has a utility representation

$$U_0(A) = \max_{a \in A, x \in a} \left[u(x) - \max_{y \in a}(v(y) - v(x)) - \max_{b \in A}(V(b) - V(a)) \right]$$

for some $u, v \in \mathcal{U}$ and $V \in \mathcal{U}_1$. Moreover, if $\succeq$ satisfies Self-Control, then the triple (u, v, V) in this representation is unique up to a positive linear transformation. The utility function U_0 has the required form (4)-(5), where

$$U(a) = \max_{x \in a}[u(x) - \max_{y \in a}(v(y) - v(x))]$$

for all menus $a \in \mathcal{M}_1$. Let $w = u + v$ and $W = U + V$. Then

$$U_0(A) = \max_{a \in A} W(a) - \max_{b \in A} V(b) \tag{13}$$

$$U(a) = w(a) - v(a) \tag{14}$$

for all $A \in \mathcal{M}_0$ and $a \in \mathcal{M}_1$.

Turn to Theorems 2 and 3. Suppose that $\succeq$ satisfies Axioms 1–5 and Binary Self-Deception (BSD for short). By Theorem 1, $\succeq$ is represented by (13). By Self-Control, there are $x^*, y^* \in X$ such that

$$\{\{x^*\}\} \succ \{\{x^*, y^*\}\} \succ \{\{y^*\}\}.$$

Then $w(x^*) > w(y^*)$ and $v(y^*) > v(x^*)$, and hence, w and v are not redundant. Without loss in generality, assume that

$$u(x^*) = v(x^*) = w(x^*) = V(\{x^*\}) = 0. \tag{15}$$

The following two lemmas obtain the required form for V.

Lemma 6. *There are $\kappa, \rho, \mu \in \mathbb{R}$ such that for all $a \in \mathcal{M}_1$,*

$$V(a) = \kappa w(a) + \rho v(a) + \mu u(a). \tag{16}$$

Proof. We claim first that for all $a, b \in \mathcal{M}_1$,

$$w(a) = w(b),\ v(a) = v(b),\ u(a) = u(b) \quad \Rightarrow \quad V(a) = V(b). \tag{17}$$

Show this claim by contradiction. Consider any $a, b \in \mathcal{M}_1$ such that $w(a) = w(b)$, $v(a) = v(b)$, $u(a) = u(b)$, but $V(b) > V(a)$. By (14), $U(a) = U(b)$ and hence, $W(b) > W(a)$. As W is continuous, then there is $\varepsilon > 0$ such that

$$W(\varepsilon\{y^*\} + (1-\varepsilon)b) > W(\varepsilon\{x^*\} + (1-\varepsilon)a).$$

Let $a^* = \varepsilon\{x^*\} + (1-\varepsilon)a$ and $b^* = \varepsilon\{y^*\} + (1-\varepsilon)b$. As $w(x^*) > w(y^*)$, $v(y^*) > v(x^*)$, and $u(x^*) > u(y^*)$, then by linearity, $w(a^*) = w(a^* \cup b^*) > w(b^*)$, $v(b^*) = v(a^* \cup b^*) > v(a^*)$, and $u(a^*) = u(a^* \cup b^*) > u(b^*)$. By (14),

$$U(a^*) > U(a^* \cup b^*) > U(b^*).$$

As $W(b^*) > W(a^*)$, then there are two possible cases.

- $W(b^*) > W(a^* \cup b^*)$. Then $V(b^*) > V(a^* \cup b^*)$. By (13), $\{b^*\} \sim \{b^*, a^* \cup b^*\}$, which contradicts BSD.
- $W(a^* \cup b^*) > W(a^*)$. Then $V(a^* \cup b^*) > V(a^*)$. By (13), $\{a^*\} \succ \{a^*, a^* \cup b^*\}$, which contradicts BSD because $u(a^*) > u(b^*)$.

This contradiction shows (17).

Take any four menus $a_1, a_2, a_3, a_4 \in \mathcal{M}_1$. Let $a = \cup_{i=1}^4 a_i$. There is i such that $w(a) \geq w(a_i)$, $v(a) \geq v(a_i)$, and $u(a) \geq u(a_i)$. Let $b = \cup_{j \neq i} a_j$. Then $w(a) = w(b)$, $v(a) = v(b)$, and $u(a) = u(b)$. By (17), $V(a) = V(b)$. Kopylov [14, Theorem 2.1] implies that the ranking that V represents on $\mathcal{M}_1$ is represented also by

$$V'(a) = \sum_{i=1}^{S} \gamma_i u_i(a) \tag{18}$$

such that $S \leq 3$, $\gamma_1, \ldots, \gamma_S \in \{-1, 1\}$, and $u_1, \ldots, u_S \in \mathcal{U}$ are not redundant. As both V' and V are linear, then without loss in generality, $V' = V$.

We claim that for all $i \in \{1, \ldots, S\}$,

$$u_i \in \mathcal{T}(w) \cup \mathcal{T}(v) \cup \mathcal{T}(u). \tag{19}$$

Wlog let $i = 1$, and suppose that $u_1 \notin \mathcal{T}(w) \cup \mathcal{T}(v) \cup \mathcal{T}(u)$. Then by Lemma 5, there are $x_1, \ldots, x_S, x_{S+1}, x_{S+2}, x_{S+3}$ such that

$$\begin{aligned}
&u_1(x_1) > u_1(x_j) \quad \text{for all } j \neq 1 \\
&u_i(x_i) \geq u_i(x_j) \quad \text{for all } i > 1 \text{ and } j \neq i \\
&w(x_{S+1}) \geq w(x_j) \quad \text{for all } j \neq S+1 \\
&v(x_{S+2}) \geq v(x_j) \quad \text{for all } j \neq S+2 \\
&u(x_{S+3}) \geq u(x_j) \quad \text{for all } j \neq S+3.
\end{aligned}$$

Let $a = \{x_1, \ldots, x_{S+3}\}$ and $b = \{x_2, \ldots, x_{S+3}\}$. Then $u_1(a) = u_1(x_1) > u_1(b)$, but $w(a) = w(b)$, $v(a) = v(b)$, $u(a) = u(b)$, and $u_j(a) = u_j(b)$ for all $j \neq 1$. Thus,

$$V(b) - V(a) = V'(b) - V'(a) = \gamma_i(u_i(x_i) - u_i(a)) \neq 0,$$

which contradicts (17).

The equalities (18) and (19) and the normalization (15) imply (16). □

The previous lemma implies that

$$\begin{aligned}
W(a) = U(a) + V(a) &= (\kappa+1)w(a) + (\rho-1)v(a) + \mu u(a) \\
&= (\kappa+1)U(a) + (\kappa+\rho)v(a) + \mu u(a)
\end{aligned} \tag{20}$$

for all menus $a \in \mathcal{M}_1$.

Lemma 7. *The functions u, v are independent. The parameters κ, ρ, μ are unique and satisfy $\rho \leq 0$, $\kappa + \rho > 0$, $\mu \geq 0$. If $\succeq$ satisfies Self-Deception, then $\mu = 0$.*

Proof. Let $a = \{x^*\}$ and $b = \{y^*\}$. By (14), $\{a\} \succ \{a \cup b\} \succ \{b\}$. By BSD, $V(a) \geq V(a \cup b)$. By (16),

$$V(a \cup b) - V(a) = \kappa(w(x^*) - w(x^*)) + \rho(v(y^*) - v(x^*)) + \mu(u(x^*) - u(x^*)) \leq 0.$$

Thus, $\rho \leq 0$. To prove the other claims of the lemma, consider two cases.

Case 1. w, v, u are redundant. Then u must be a positive linear transformation of w or v. If $u = \alpha v$ for some $\alpha > 0$, then $w = u + v$ and v are redundant. Thus, $u = \alpha w$ for some $\alpha > 0$. Then $v = (\alpha - 1)w$. As w and v are not redundant, then $\alpha \in (0, 1)$. For all $a \in \mathcal{M}_1$,

$$\begin{aligned}
V(a) &= \kappa w(a) + \rho v(a) + \mu u(a) = (\kappa' + \rho)U(a) + \rho v(a), \\
W(a) &= (\kappa' + 1)U(a) + (\kappa' + \rho)v(a),
\end{aligned}$$

where $\kappa' = \kappa + \mu\alpha$. Suppose that $\kappa' + \rho < 0$. Take $\alpha, \beta \in (0, \frac{1}{2})$ such that

$$1 < \frac{\alpha}{\beta} \frac{w(x^*) - w(y^*)}{v(y^*) - v(x^*)} < 1 + \left| \frac{\kappa' + \rho}{\kappa' + 1} \right|. \tag{21}$$

Let $a = \{x^*, y^*\}$ and $b = \{\alpha y^* + (1-\alpha)x^*, \beta x^* + (1-\beta)y^*\}$. Then

$$\begin{aligned} w(a \cup b) - w(b) &= w(x^*) - w(\alpha y^* + (1-\alpha)x^*) = \alpha(w(x^*) - w(y^*)) > 0 \\ v(a \cup b) - v(b) &= v(y^*) - v(\beta x^* + (1-\beta)y^*) = \beta(v(y^*) - v(x^*)) > 0. \end{aligned}$$

By (21) and (20),

$$\begin{gathered} U(a \cup b) - U(b) = \alpha(w(x^*) - w(y^*)) - \beta(v(y^*) - v(x^*)) > 0 \\ |(\kappa' + 1)(U(a \cup b) - U(b))| < |\kappa' + \rho| \, \beta(v(y^*) - v(x^*)) \\ W(a \cup b) - W(b) = (\kappa' + 1)(U(a \cup b) - U(b)) + (\kappa' + \rho)\beta(v(y^*) - v(x^*)) < 0. \end{gathered}$$

Therefore, $\{a \cup b\} \succ \{b\}$, but $\{b, a \cup b\} \sim \{b\}$, which contradicts BSD. Thus, $\kappa' + \rho \geq 0$. Thus, for all $x \in X$,

$$V(\{x\}) = (\kappa' + \rho)U(\{x\}) + \rho\tfrac{\alpha-1}{\alpha}u(x) = \gamma U(\{x\}),$$

where $\gamma = (\kappa' + \rho) + \rho\frac{\alpha-1}{\alpha}$ is positive. By (13) and (14), for all $x, y \in X$,

$$\{\{x\}\} \succeq \{\{y\}\} \quad \Rightarrow \quad \{\{x\}\} \sim \{\{x\}, \{y\}\} \succeq \{\{y\}\},$$

which violates Self-Control.

Case 2. w, v, u are not redundant. Then u and v are independent, and there are $x, y, z \in X$ such that

$$\begin{aligned} &w(x) > w(y) \vee w(z) \\ &v(y) > v(x) \vee v(z) \\ &u(z) > u(x) \vee u(y). \end{aligned}$$

Suppose that $\mu < 0$. Take $\alpha \in (0, 1)$ such that

$$\alpha(\kappa + 1)(w(x) - w(y)) + \mu(u(z) - u(x)) < 0.$$

Let $a = \{x, y, z\}$ and $b = \{y, \alpha y + (1-\alpha)x\}$. Then

$$\begin{aligned} w(a \cup b) - w(b) &= w(x) - w(\alpha y + (1-\alpha)x) = \alpha(w(x) - w(y)) > 0 \\ v(a \cup b) - v(b) &= v(y) - v(y) = 0 \\ u(a \cup b) - u(b) &= u(z) - u(\alpha y + (1-\alpha)x) \geq u(z) - u(x) > 0. \end{aligned}$$

By (14), $U(a \cup b) - U(b) = w(a \cup b) - w(b) > 0$. By (20),

$$W(a \cup b) - W(b) = (\kappa + 1)(w(a \cup b) - w(b)) + \mu(u(a \cup b) - u(b)) \leq \\ \alpha(\kappa + 1)(w(x) - w(y)) + \mu(u(z) - u(x)) < 0.$$

Therefore, $\{a \cup b\} \succ \{b\}$ and $\{b, a \cup b\} \sim \{b\}$. These rankings violate BSD. Thus, $\mu \geq 0$.

Suppose that $\kappa + \rho < 0$. Take $\alpha, \beta \in (0, \frac{1}{2})$ such that

$$1 < \frac{\alpha}{\beta} \frac{w(x) - w(y)}{v(y) - v(x)} < 1 + \left| \frac{\kappa + \rho}{\kappa + 1} \right|. \tag{22}$$

Let $a = \{x, y, z\}$ and $b = \{\alpha y + (1 - \alpha)x, \beta x + (1 - \beta)y, z\}$. Then

$$w(a \cup b) - w(b) = w(x) - w(\alpha y + (1 - \alpha)x) = \alpha(w(x) - w(y)) > 0$$
$$v(a \cup b) - v(b) = v(y) - v(\beta x + (1 - \beta)y) = \beta(v(y) - v(x)) > 0$$
$$u(a \cup b) - u(b) = u(z) - u(z) = 0.$$

By (22) and (20),

$$U(a \cup b) - U(b) = \alpha(w(x) - w(y)) - \beta(v(y) - v(x)) > 0$$
$$|(\kappa + 1)(U(a \cup b) - U(b))| < |\kappa + \rho| \, \beta(v(y) - v(x))$$
$$W(a \cup b) - W(b) = (\kappa + 1)(U(a \cup b) - U(b)) + (\kappa + \rho)\beta(v(y) - v(x)) < 0.$$

Therefore, $\{a \cup b\} \succ \{b\}$, but $\{b, a \cup b\} \sim \{b\}$, which contradicts BSD. Thus, $\kappa + \rho \geq 0$.

Suppose that $\kappa + \rho = 0$. Then for all $x' \in X$,

$$V(\{x'\}) = \kappa U(\{x'\}) + \mu u(x') = (\kappa + \mu)U(\{x'\}),$$

where $\kappa + \mu = -\rho + \mu \geq 0$. This equality contradicts Self-Control (see the proof of Case 1.) Thus, $\kappa + \rho > 0$.

Suppose that $\succeq$ satisfies Self-Deception. Let $a = \{x, y\}$ and $b = \{y, z\}$. Then $\{a \cup b\} \succ \{b\}$. By Self-Deception,

$$V(a) - V(a \cup b) = \mu(u(x) - u(z)) \geq 0.$$

Thus, $\mu = 0$.

Finally, note that

$$\kappa = \frac{V(\{x,y,z\} - V(\{\alpha z + (1-\alpha)x, y, z\})}{\alpha(w(x) - w(z))}$$
$$\rho = \frac{V(\{x,y,z\} - V(\{x, \alpha x + (1-\alpha)y, z\})}{\alpha(v(y) - v(x))}$$
$$\mu = \frac{V(\{x,y,z\} - V(\{x, y, \alpha x + (1-\alpha)z\})}{\alpha(u(z) - u(x))}$$

for all sufficiently small α. These equations show that all of these parameters are unique. □

Lemmas 6 and 7 imply that V has the required form

$$V(a) = \kappa U(a) + \lambda v(a) + \mu u(a) = \kappa w(a) + (\lambda - \kappa)v(a) + \mu u(a), \qquad (23)$$

where $\kappa \geq \lambda = \kappa + \rho > 0$, $\mu \geq 0$, and $u, v \in \mathcal{U}$ are independent.

Conversely, suppose that $\succeq$ has representation (13), (14), and (23). Take any $a, b \in \mathcal{M}_1$ such that $\{a \cup b\} \succ \{b\}$. By (23),

$$V(a \cup b) - V(b) = \kappa(U(a \cup b) - U(b)) + \lambda(v(a \cup b) - v(b)) + \mu(u(a \cup b) - u(b)) \geq 0$$

because $U(a \cup b) > U(b)$, $v(a \cup b) \geq v(b)$, and $u(a \cup b) \geq u(b)$. By (13),

$$\{a \cup b\} \sim \{b, a \cup b\} \succ \{b\}.$$

Moreover,

$$V(a) - V(a \cup b) = \kappa(w(a) - w(a \cup b)) + (\lambda - \kappa)(v(a) - v(a \cup b)) + \mu(u(a) - u(a \cup b)) \geq \mu(u(a) - u(a \cup b))$$

because $w(a) = w(a \cup b)$, $v(a \cup b) \geq v(a)$, and $\lambda - \kappa \leq 0$. Therefore, the ranking $\{a\} \succ \{a, a \cup b\}$ implies that $\mu > 0$ and $u(b) > u(a)$. Thus $\succeq$ satisfies Binary Self-Deception, and if $\mu = 0$, then $\succeq$ satisfies Self-Deception.

As u and v are independent, then w and v are not redundant. Take $x, y \in X$ such that $w(x) > w(y)$ and $v(y) > v(x)$. By (14),

$$\{\{x\}\} \succ \{\{x, y\}\} \succ \{\{y\}\}.$$

Let $v' = (\kappa + \mu)u + \lambda v$ and $w' = u + v'$. As u and v are independent and $\lambda > 0$, then the functions w' and v' are not redundant. Take $x', y' \in X$ such that $w'(x') > w'(y')$ and $v'(y') > v'(x')$. By (13),

$$\{\{x'\}\} \succ \{\{x'\}, \{y'\}\} \succ \{\{y'\}\}.$$

Thus, $\succeq$ satisfies Self-Control.

Turn to Theorem 4. Suppose that $\succeq$ and $\succeq^*$ have representations (4)-(6) with components $(u, v, \kappa, \lambda, \mu)$ and $(u, v, \kappa^*, \lambda^*, \mu^*)$ such that $\kappa \geq \lambda > 0$, $\kappa^* \geq \lambda^* > 0$, and $\mu, \mu^* \geq 0$.

Suppose that $\frac{\lambda^*}{\kappa^*} \geq \frac{\lambda}{\kappa}$ and $\frac{\mu^*}{\lambda^*} = \frac{\mu}{\lambda}$. Then $U = U^*$ and for all $a, b \in \mathcal{M}_1$,

$$\begin{aligned}
&\{a\} \succ \{a, b\} \;\Rightarrow\; U(a) > U(b) \;\text{ and }\; V(b) > V(a) \;\Rightarrow \\
&[U(b) - U(a)] + \tfrac{\lambda}{\kappa}[v(b) - v(a) + \tfrac{\mu}{\lambda}u(b) - \tfrac{\mu}{\lambda}u(a)] > 0 > U(b) - U(a) \;\Rightarrow \\
&[U^*(b) - U^*(a)] + \tfrac{\lambda^*}{\kappa^*}[v(b) - v(a) + \tfrac{\mu^*}{\lambda^*}u(b) - \tfrac{\mu^*}{\lambda^*}u(a)] > 0 > U^*(b) - U^*(a) \;\Rightarrow \\
&U^*(a) > U^*(b) \;\text{ and }\; V^*(b) > V^*(a) \;\Rightarrow\; \{a\} \succ^* \{a, b\}.
\end{aligned}$$

Thus, $\succeq^*$ is more self-deceptive than $\succeq$.

Conversely, suppose that $\succeq^*$ is more self-deceptive than $\succeq$. As u and v are independent, then the functions u, v, and $w = u + v$ are not redundant. By Lemma 5, there are $x, y, z \in X$ such that

$$\begin{aligned}
&w(x) > w(y) \vee w(z) \\
&v(y) > v(x) \vee v(z) \\
&u(z) > u(x) \vee u(y).
\end{aligned}$$

As $w, v, u \in \mathcal{U}$, then for any $\alpha, \gamma > 0$, there exist $x', y', z' \in X$ such that

$$\begin{aligned}
&w(x) > w(x') > w(y) \vee w(y') \vee w(z) \vee w(z') \\
&v(y) > v(y') > v(x) \vee v(x') \vee v(z) \vee v(z') \\
&u(z) > u(z') > u(x) \vee u(x') \vee u(y) \vee u(y') \\
&\frac{w(x) - w(x')}{v(y) - v(y')} = \alpha \\
&\frac{v(y) - v(y')}{u(z) - u(z')} = \gamma.
\end{aligned} \tag{24}$$

Show the inequalities $\frac{\lambda^*}{\kappa^*} \geq \frac{\lambda}{\kappa}$ and $\frac{\mu^*}{\lambda^*} = \frac{\mu}{\lambda}$ by contradiction. Consider three cases.

Case 1. $\frac{\mu^*}{\lambda^*} > \frac{\mu}{\lambda}$. Take γ such that $\frac{\mu^*}{\lambda^*} > \gamma > \frac{\mu}{\lambda}$ and α such that $1 > \alpha > 1 - \frac{\gamma\lambda - \mu}{\gamma\kappa}$. Take $x', y', z' \in X$ that satisfy (24). Let $a = \{x', y', z\}$ and $b = \{x, y, z'\}$. Then

$$U(a) - U(b) = (w(x') - v(y')) - (w(x) - v(y)) = (1 - \alpha)(v(y) - v(y')) > 0$$
$$V(b) - V(a) = \left(-\kappa(1 - \alpha) + \lambda - \tfrac{\mu}{\gamma}\right)(v(y) - v(y')) > 0$$
$$V^*(b) - V^*(a) = \left(-\kappa^*(1 - \alpha) + \lambda^* - \tfrac{\mu^*}{\gamma^*}\right)(v(y) - v(y')) < 0$$

because

$$-\kappa(1 - \alpha) + \lambda - \tfrac{\mu}{\gamma} > 0 > -\kappa^*(1 - \alpha) + \lambda^* - \tfrac{\mu^*}{\gamma}.$$

Thus, $\{a\} \succ \{a, b\}$, but $\{a\} \sim^* \{a, b\}$, which contradicts the assumption that $\succeq^*$ is more self-deceptive than $\succeq$.

Case 2. $\frac{\mu^*}{\lambda^*} < \frac{\mu}{\lambda}$. Take γ such that $\frac{\mu^*}{\lambda^*} < \gamma < \frac{\mu}{\lambda}$ and α such that $1 < \alpha < 1 + \frac{\mu - \gamma\lambda}{\gamma\kappa}$. Take $x', y', z' \in X$ that satisfy (24). Let $a = \{x, y, z'\}$ and $b = \{x', y', z\}$. Then

$$U(a) - U(b) = (w(x) - v(y)) - (w(x') - v(y')) = (\alpha - 1)(v(y) - v(y')) > 0$$
$$V(b) - V(a) = \left(-\kappa(\alpha - 1) - \lambda + \tfrac{\mu}{\gamma}\right)(v(y) - v(y')) > 0$$
$$V^*(b) - V^*(a) = \left(-\kappa^*(\alpha - 1) - \lambda^* + \tfrac{\mu^*}{\gamma^*}\right)(v(y) - v(y')) < 0$$

because

$$-\kappa(\alpha - 1) - \lambda + \tfrac{\mu}{\gamma} > 0 > -\kappa^*(\alpha - 1) - \lambda^* + \tfrac{\mu^*}{\gamma}.$$

Thus, $\{a\} \succ \{a, b\}$, but $\{a\} \sim^* \{a, b\}$, which contradicts the assumption that $\succeq^*$ is more self-deceptive than $\succeq$.

Case 3. $\frac{\mu^*}{\lambda^*} = \frac{\mu}{\lambda}$ and $\frac{\lambda^*}{\kappa^*} < \frac{\lambda}{\kappa}$. Take α such that $1 - \frac{\lambda^*}{\kappa^*} > \alpha > 1 - \frac{\lambda}{\kappa}$. Take $x', y' \in X$ that satisfy (24). (z' is not required here.) Let $a = \{x', y', z\}$ and $b = \{x, y, z\}$. Then

$$U(a) - U(b) = (w(x') - v(y')) - (w(x) - v(y)) = (1 - \alpha)(v(y) - v(y')) > 0$$
$$V(b) - V(a) = (-\kappa(1 - \alpha) + \lambda)(v(y) - v(y')) > 0$$
$$V^*(b) - V^*(a) = (-\kappa^*(1 - \alpha) + \lambda^*)(v(y) - v(y')) < 0$$

because

$$-\kappa(1 - \alpha) + \lambda > 0 > -\kappa^*(1 - \alpha) + \lambda^*.$$

Thus, $\{a\} \succ \{a,b\}$, but $\{a\} \sim^* \{a,b\}$, which contradicts the assumption that $\succeq^*$ is more self-deceptive than $\succeq$.

Thus, $\frac{\mu^*}{\lambda^*} = \frac{\mu}{\lambda}$ and $\frac{\lambda^*}{\kappa^*} \geq \frac{\lambda}{\kappa}$.

References

[1] C. Aliprantis and K. Border. *Infinite Dimensional Analysis*. Springer, 1999.

[2] A. Barnes. *Seeing Through Self-Deception*. Cambridge University Press, New York, 1997.

[3] R. Baumeister, H. T., and T. D. *Losing Control: How and Why People Fail at Self-Regulation*. Academic Press, San Diego, CA, 1994.

[4] R. Benabou and J. Tirole. Self-confidence and personal motivation. *Quarterly Journal of Economics*, 135:871–915, 2002.

[5] J. Bermudez. Self-deception, intentions and contradictory beliefs. *Analysis*, 60(4):309–319, 2000.

[6] M. Brunnermeier and P. J. Optimal expectations. *American Economic Review*, 95(4):1092–1118, 2005.

[7] E. Dekel, B. L. Lipman, and A. Rustichini. Representing preferences with a unique subjective state space. *Econometrica*, 69:891–934, 2001.

[8] E. Dekel, B. L. Lipman, and A. Rustichini. Temptation-driven preferences. *Review of Economic Studies*, 2009. forthcoming.

[9] L. Epstein. An axiomatic model of non-Bayesian updating. *Review of Economic Studies*, 73:413–436, 2006.

[10] H. Fingarette. *Self-Deception*. UC California Press, Berkeley, 1969, 2000.

[11] F. Gul and W. Pesendorfer. Temptation and self-control. *Econometrica*, 69:1403–1435, 2001.

[12] R. Gur and H. Sackeim. Self-deception: A concept in search of a phenomenon. *Journal of Personality and Social Psychology*, 37(2):147–169, 1979.

[13] I. Kopylov. Perfectionism and choice. Mimeo, UC Irvine, 2008.

[14] I. Kopylov. Finite additive utility representations for preferences over menus. *Journal of Economic Theory*, 144:354–374, 2009.

[15] I. Kopylov. Temptations in general settings. Mimeo, UC Irvine, 2009.

[16] Z. Kunda. The case for motivated reasoning. *Psychological Bulletin*, 108(3):480–498, 1990.

[17] N. Levy. Self-deception and moral responsibility. *Ratio (new series)*, 17:294–311, 2004.

[18] A. Ludwig. *Understanding the Alcoholic's Mind: The Nature of Craving and How to Control It*. Oxford University Press, New York, 1988.

[19] A. Mele. *Self-Deception Unmasked*. Princeton University Press, Princeton, 2001.

[20] J. Noor. Commitment and self-control. *Journal of Economic Theory*, 135:1–34, 2007.

[21] J. Noor. Temptation, welfare, and revealed preference. Mimeo, Boston University, 2009.

[22] J.-P. Sartre. *Being and Nothingness*. New York, 1956.

[23] B. Szabados. Collaborative companions: The relationship of self-deception and excuse-making. In M. M., editor, *Self-Deception and Self-Understanding: New Essays in Philosophy and Psychology*. University Press of Kansas, Lawrence, 1985.

[24] W. Talbott. Intentional self-deception in a single coherent self. *Philosophy and Phenomenological Research*, 55:27–74, 1995.

[25] S. Taylor. *Positive Illusions: Creative Self-Deception and the Healthy Mind*. Basic Books, New York, 1989.

[26] S. Taylor and J. Brown. Wishful thinking and self-deception. *Analysis*, 33(6):201–205, 1973.

[27] S. Taylor and J. Brown. Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*, 103:193–210, 1988.